Zeitschrift: Comtec : Informations- und Telekommunikationstechnologie = information and telecommunication technology

Herausgeber: Swisscom

Band: 76 (1998)

Heft: 6

Artikel: Voice service opportunities : spoken language processing in Telecom services

Autor: Cochard, Jean-Luc

DOI: https://doi.org/10.5169/seals-877307

From the Exploration Programmes of Corporate Technology (1)

Voice Service Opportunities

Spoken Language Processing in Telecom Services

The goal of this paper is to give a general overview of the activities that have been planned within the Corporate Technology Exploration Program EP9701: Voice Services Opportunities. Since one important part of this program is dedicated to Automatic Spoken Language Processing, we consider it useful to start with a basic presentation of what is meant by this notion. The central part of this paper consists of a description of the structure we adopted for the projects to be run within the EP9701. Besides the fact that research topics have to be investigated to gain extensive know-how in spoken language processing and IP telephony, it was considered a major issue to generate a strong team-working spirit.

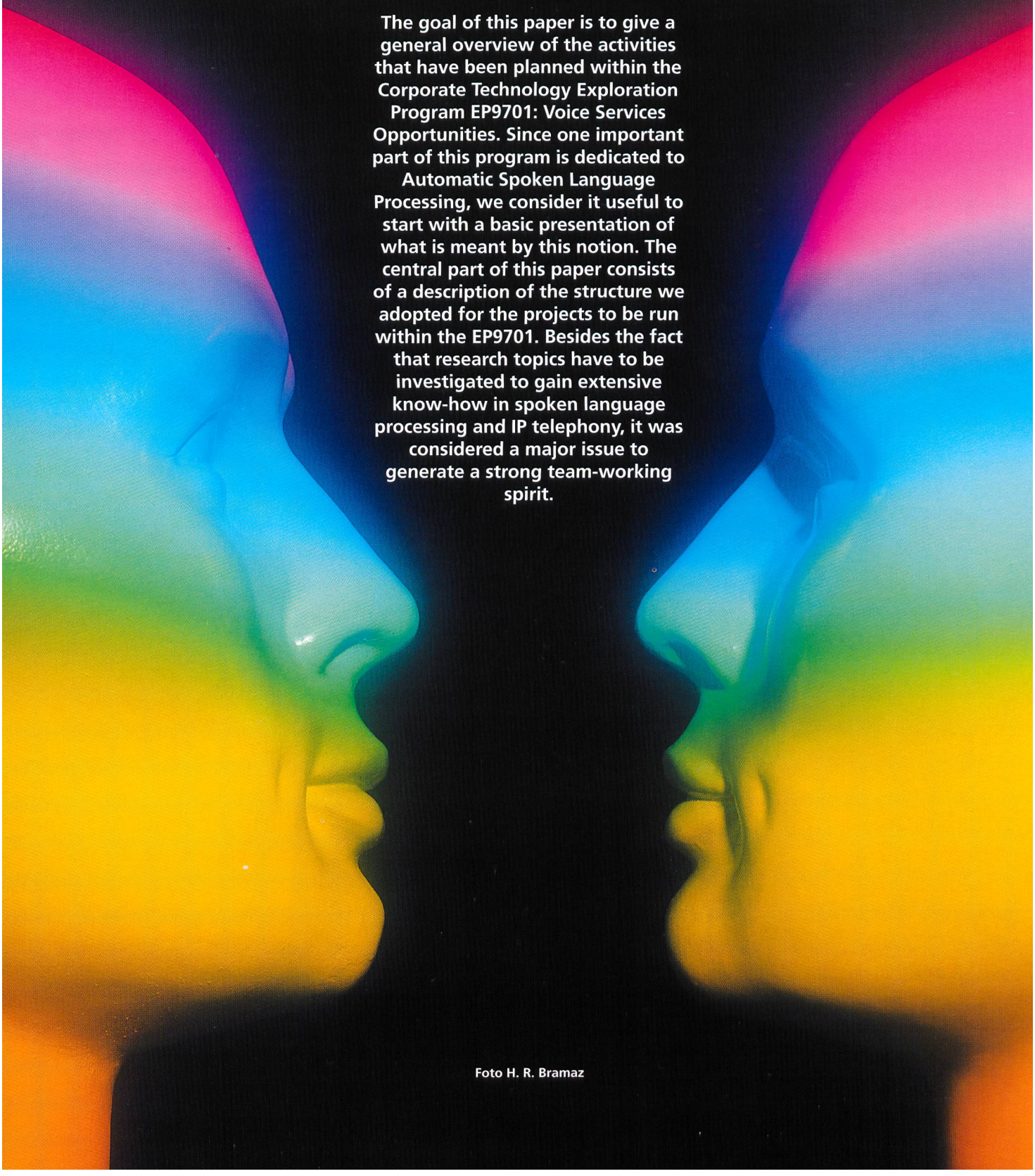

Foto H. R. Bramaz

It was also decided to seize the opportunity to include in this paper a presentation of the academic and private research institutes working in collaboration with us for this EP9701. They were asked to describe briefly their dedicated

JEAN-LUC COCHARD, BERN

research topics and to highlight some of the challenges they are currently taking up. We hope that both descriptions: application driven, on the one hand, with the concrete objectives of the EP9701, and more basic research oriented, on the other, with the speech labs presentations, will be useful to understand how complex the problems still are and what is currently available on the market to be shortly introduced into advanced communication services.

What is meant by "Automatic Spoken Language Processing"?

The automatic processing of spoken language utterances has been a constantly investigated research field for more than 40 years in academic laboratories as well as in the industry. To implement a software able to understand spoken commands was a goal for researchers in the early age of computing. The availability of fast, cheap and reliable processors, the mastering of complex algorithms and the sophisticated modeling of speech units are the key elements that gave rise to many commercial software packages allowing an end-user to speak to a computer. It is thus possible to dictate a text or to command the graphical user interface of a PC by voice. At the other end of a communication process, text-to-speech technology allows the reading of a given ASCII text.

To limit the scope of this introduction, automatic spoken language processing will be considered from now on only in the framework of telecom applications. To give an overview of what a typical advanced communication service could consist of, three distinct components have to be described, each of them having its own problematic:

- A *speech recognition* component allows a user to talk to an automatic system via his/her telephone. In this context (Fig. 1), speech is a command and control means, and the purpose of the application is to deliver spoken information. The MessageBox service from Swisscom could be such an application, except that in the current version, the user cannot speak to his/her mailbox yet but interacts by pressing sequences of touch-tone keys. There exists a wide range of applications for which speech recognition technology can greatly improve user-friendliness and expressive power of the current touch-tone menu interface. For instance, a speech activated system is not limited by a predefined number of command words – with touch-tone applications, a maximum of 12 functions can be offered in parallel – and the commands can be more compact and intuitive, such as for example «Delete message number one».
- A *speaker verification* component allows secure access to confidential information using voice discriminative characteristics. Voice conveys a lot of non linguistic information. For example, if someone is talking to you, you can say if the speaker is tired, stressed, out of breath, or if he/she has a cold, etc. A more useful information for our purposes is that when listening carefully to someone without seeing them, you get enough information to discriminate their voice from others and thus assign an identity to it. In this context, there may be two distinct problems. One is to guess the identity of someone who doesn't introduce him/herself. For an automatic system, this is known as a *speaker identification* task. The second problem is to get some confidence on the identity of someone who is claiming his/her identity. An automatic system implementing this function is performing a *speaker verification task* (Fig. 2). For telematics services, only speaker verification is useful, and therefore it is addressed in the EP9701.
- A *speech synthesis* component is used to provide information through the telephone channel. There exist many technical solutions to reach this objective. The simpler one consists in playback of prerecorded messages. A more sophisticated and versatile approach is to use a concatenation mechanism to connect small prerecorded speech units. Thus the system is able to generate a spoken version of virtually any written text. The latter approach is unquestionably more flexible but it does not guarantee a result as natural as if messages were entirely recorded by a human speaker.

The Exploration Program EP9701

Besides sparkling multimedia facilities, Internet has set itself as the de facto standard of the present information society paradigm. One less visible by-product of this technological development is the advent of a "packet-switched" network, while the current telephony network is "circuit-switched". As far as voice communication is concerned, this represents a clear network paradigm shift. The emerging Voice over IP (Internet Protocol) technology is one of its instances.
In the area of user interfaces, automatic spoken language applications start to mimic human-like operator interactions. Significant progress has been achieved in spoken dialogue systems, speech recognition, speaker verification, speech synthesis, and ultimately, machine translation of spoken dialogues.
EP9701 will focus on packet-switched telephony using enhanced spoken lan-

Besides sparkling multimedia facilities, Internet has set itself as the de facto standard of the present information society paradigm. One less visible by-product of this technological development is the advent of a "packet-switched" network, while the current telephony network is "circuit-switched". As far as voice communication is concerned, this represents a clear network paradigm shift. The emerging Voice over IP (Internet Protocol) technology is one of its instances.
In the area of user interfaces, automatic spoken language applications start to mimic human-like operator interactions. Significant progress has been achieved in spoken dialogue systems, speech recognition, speaker verification, speech synthesis, and ultimately, machine translation of spoken dialogues.

EP9701 "Voice Service Opportunities" will focus on packet-switched telephony using enhanced spoken language processing technologies. These new communication means are expected to have a dramatic impact on all voice services.

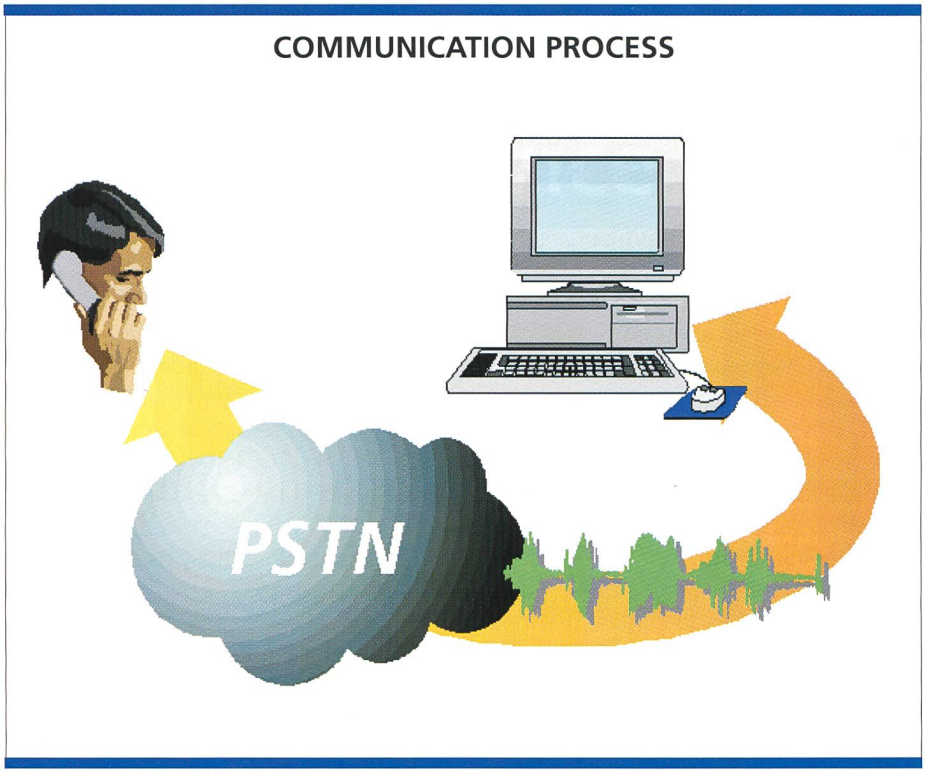

Figure 1. Illustration of a communication process between a user and an automatic service. The user is issuing commands by voice using a telephone handset in order to get information stored on a computer which symbolizes the hardware running the automatic service. Information is then delivered through the same voice channel; it can be either play-back of prerecorded speech or execution of a text-to-speech program.

guage processing technologies. These new communication means are expected to have a dramatic impact on all voice services.

Overall Structure and Underlying Concepts

During the kick-off meeting of EP9701 that took place in the beginning of January 1998, it was decided to reorganise the various activities that were mentioned in the Funding Request into a matrix structure, in order to strengthen teamwork. The columns of this two-dimensional matrix (Fig. 3) represents the technological activities, namely Voice over IP, text-to-speech (TTS), speaker verification (SV) and speech recognition (SR) research themes. For all these topics of interest, our goal is to acquire or improve our know-how. This means that the project leader of each of these projects has to be aware of new developments, to understand the major characteristics of commercially available products, and finally to conduct studies on a limited number of topics within these large thematic fields.

On the horizontal dimension, two general topics strongly related to future services are addressed. The purpose of these two projects is not really to implement new services but instead to conduct in-depth experiments on the best possible integration of the above mentioned technologies in coherent, user-friendly but also futuristic telematics applications. Work done along this service-oriented dimension will be essentially an iterative process based on cycles including the following steps (Fig. 4):

- *Brainstorming* is the initial step involving possibly all collaborators of the EP9701. Its main objective is to give an opportunity to "dream" in terms of what could be a sophisticated futuristic service taking advantage of all the technologies the program is able to provide. One aspect leading our discussions is not necessarily to invent a completely new service from scratch, but rather to consider the possibility of merging existing, separate, value-added services in one coherent and uniform, thus user-friendly, more advanced prototype.
- Results of the brainstorming are translated into a set of *Requirements;* some are the goals to be achieved by technology-oriented projects, and others are addressing integration matters assigned to the service-oriented projects.
- *Design* of the application has to precisely describe the interactions of its various components. Since this class of applications is intrinsically event-driven, one of the best possible descriptions is a flowchart-based model using graphical representations. It is expected that the design outcome could also be used to communicate our objectives to our business units (BUs).
- *Implementation* refers to the integration of the various components provided by the technology-oriented projects and the development of the application specific layer. For example, the complete dialogue and all the functions specific to the prototype under development have to be set up in this particular step.
- *Testing* is a key issue in this iterative prototyping process. However, we consider ourselves to be in the worst position to evaluate our work, especially the quality of interaction offered by a new service prototype making extensive use of spoken language processing capabilities. Therefore, we plan to delegate this task to a limited set of targeted end-users and to BUs representatives. We expect valuable feedback from beta-tests involving laymen.
- *Evaluation* will highlight the weaknesses and design errors that have been encountered by end-users while using the prototype. Integrating strategic input from BUs will also be one major concern during this step. Therefore, taking into account the content of this review will be the objective of a next iteration of the prototype generation process. Another outcome of this step should be the definition of concrete Application Projects for business units, by reusing the most successful features of the prototype and fine tuning them in order to build a commercially available service.

In addition to this description of the general philosophy of the EP9701, the rest of this section is devoted to more specific descriptions of the individual projects either technology-oriented or service-oriented.

Voice Over IP

The market increasingly demands the availability of multimedia and multiparty services with extensive support of mobility features.

Packet-switched *voice and multimedia* combined with data compression appears to be the 'royal way' to low cost services. Data services, for which packet-switched networks were made, do not suffer from network imperfections such as delay, packet-loss, etc. (given enough time, all shortcomings of the network, excepting delay, may be corrected in the terminals). On the other hand, real-time services represent a new and difficult challenge to such networks. It is imagined that in the future, POTS will run over data networks via gateways and be integrated, together with high quality sound and video, into data services. Globe-spanning data networks, owned and managed by a single company or consortium, will be able to offer at a much reduced price the necessary QoS-guarantees that customers expect.

An important change in paradigm taking place in telecommunications today – *the passage from the circuit-switched, speech-centric world to a packet-switched, data-centric one.* On the technical level, IP and ATM are two of the most important protocols capable of integrating today's and tomorrow's telecommunications services. While IP is already present in most PCs, ATM is still struggling to reach out from the backbone towards the terminal. A certain potential for symbiosis and convergence of these technologies is evident and it is clear that *IP must acquire some ATM-like features* in order to be successful.

Objectives

This project will address technology for the ubiquitous provisioning of services in the fast evolving telecommunications market place. This will lead to issues related to the design and engineering of advanced communication services and experiments on their use. Topics to be addressed include:

- Build and make a communications system available for use by 'Personal IP Telephony' and facilities to test it.
- Develop and test tools to objectively measure quality (for IP-Telephony only to start with).
- In-depth evaluation of lab system to identify areas (codec/IP-Network/HW-SW) where significant improvements can be made.
- Organizing beta-tests to evaluate perceived (customers' reactions to new services) and real transmission quality.
- Audio & Video over IP for a multimedia Web-presentation[1] of EP9701.

Speaker Verification

The ongoing development of telematics services to access or change private data shows the need for automated and user-friendly authentication methods. From the point of view of the service providers, telematics transactions provide enormous savings, due to the increased efficiency of personnel, savings in office space, and so on. From the viewpoint of the end-user, telematics transaction services will increase the accessibility of a wide range of services, because access to them is not limited to conventional office hours. Voice-based access will make transaction services available from the type of terminal that is presently most ubiquitous (and which will remain in that position for a very long time to come) – the standard telephone handset. The voice telephone is likely to be the single most important and widespread terminal in the fast-growing cellular networks. Previous research and technical development (R&TD) has shown that ease of access (in other words, user-friendliness) is as important an issue as security in transaction services. It has been shown that the combined deployment of automatic speech recognition and speaker verification can provide the required level of user-friendliness.

Objectives

The Speaker Verification project[2] intends, on one hand, to develop and test secure telematics services using caller authentication by voice, and, on the other hand, to improve the robustness of in-house available SV algorithms. This new authentication technique will be tested in terms of user acceptance and robustness of the technology to clarify its deployment in Swisscom services. The following tasks will be carried out:

- Beta-testing of two prototypes: a Voice-based Calling Card service and a Personal Call Assistant (PCA),
- Robustness evaluation of speaker models with various approaches for the enrollment procedure of new subscribers[3],
- Improvement of SV algorithms in noisy environments.

Text-to-Speech

Services which allow a telephone access to some information or a telephone-based transaction must be able to output speech to the user. This speech output capability is needed for user/service interaction (along with touch-tone recognition or automatic speech recognition to handle user input) and for actual information delivery.

Prerecorded human speech can be used when the range of text to be pronounced is easily predictable or has a limited variability (e.g. dialogue prompts, time-table information). In contrast is the situation when the information to be read is unpredictable or varies substantially. In this latter case, it becomes unrealistic to record all possible utterances from a human speaker, and a text-to-speech synthesis system can be applied at the cost of reduced speech quality. Potential applications such as an automatic reverse directory system (RDS), the automatic reading of e-mails, the retrieval of information bulletins or of Web contents often involve text of various linguistic origins. Names (person's name and address, company and product names, etc.) are typical examples of this multilingual characteristic. Several commercial TTS products are available for many different languages, but as of today, none of them permits languages to be interleaved smoothly because, linguistically and acoustically, each language is processed independently from one another.

Objectives

The project's goal is to develop a multilingual TTS system capable of reading

[1] This task, while not in the main line of the Voice over IP project, has for immediate purpose to present the EP9701 to a potentially wide public in an informative but attractive and (if possible) entertaining way. Beyond that, this project is aimed at assessing the viability of multicasting as a communication tool: With employees, partners, and customers located all over Switzerland and in dozens of locations world-wide, Swisscom's ongoing challenge is to provide its geographically distributed employees with timely, consistent information (corporate communications, training, etc). Multicasting, or live video on the desktop, represents one of today's most promising technologies for responding to diverse needs for information.

[2] A large part of these activities will be executed within the framework of a European Telematics Project called PICASSO, currently running until July 2000.

[3] To use a automatic service secured by a SV component, all subscribers will have to go through an enrolment phase essential to train the speaker authentication models.

aloud text in several individual languages (2 to 4 depending on resources) and also in a mixed-language context. Data from the telephone directory listings (ETV) will constitute the first target along the path towards multilinguality. Further steps will depend on the precise multilingual needs of applications such as reading of e-mail or of Web contents, and will be defined accordingly.
Besides this main goal, further specific developments will be aimed at facilitating the integration of the TTS module into a dialogue system:
- Software optimization and packaging (API),
- Implementation of an RDS system,
- Specific prosody generation (interrogative/exclamatory sentences for dialogue situations, spelling, reading numbers),
- Speech generation mark-up language for Web-to-speech applications.

Speech Recognition

Most interaction in voice services are limited by the input interface capability (touch-tone menus). In order to expand the potential market and to automate some of the current voice services, this project will first identify the most effective speech recognition technology available on the Swiss market.
Many parameters have to be carefully evaluated when comparing different speech recognizers, especially in a context of future operational services. However, within EP9701, one initial purpose is to concentrate on technological aspects of speech recognition over the telephone in order to increase the internal know-how. It is known that prototypes of services can already be implemented with commercially available speech recognition systems with their well-known limitations:
- Simple pronunciation modeling – French, German and Italian models are not especially tuned for Swiss-French, Swiss-German, and Swiss-Italian speakers,
- Robustness problem against channel variations – A speech recognizer with fixed network models will perform poorly with mobile telephone speech conditions.

In the academic world, many approaches are currently investigated in the direction of more robust speech recognition systems and more sophisticated pronunciation modeling. Mastering these two central problems is a key issue toward a sensible improvement of an overall system performance.

Objectives

This research project will be organized as follows:
- First, we will draw a clear picture of the state-of-the-art of commercially available speech recognition systems by running benchmarks on speech samples from Swiss speakers,
- Selection of the most promising approaches to improve state-of-the-art systems with respect to the two above mentioned problems,
- Conducting in-depth experiments on these new approaches in close relation with academic laboratories.

Ultimately, by improving the state-of-the-art, this project will clearly demonstrate the outstanding and in-house knowledge available for the business units. One expected outcome will be to enhance the robustness of speech recognition in typical voice services.

Personal IP Telephony

In the current total communication age, a principle stating that everybody should be able to reach anybody else at any time from anywhere, and conversely, that everybody should be reachable at any time and anywhere, is widely accepted in the business world. This is the primary purpose of some value-added services such as voice messaging, One Number-like, and Personal Call Assistant products to answer this market demand. In this respect, only few products include speech recognition and text-to-speech capabilities and if they do, they are very limited (e.g. "Please say 'one' for ..., say 'two' for ..., etc."). To our knowledge no product includes state-of-the-art speech recognition technology, either latest development in speaker verification, or a full multilingual text-to-speech system. In this context, the definition of a new generation of personal telephony services has to be closely tied to experiments with all the above mentioned technologies.
As was clearly stated in the presentation of the Voice over IP project, IP network can be used to compete with the standard fixed network for voice communication at lower cost. But in the context of EP9701, one major interest is to investigate the complementary characteristics of both networks. IP has proven to be a powerful infrastructure for accessing multimedia information. It could greatly improve the control and configuration procedures for customized settings within an advanced communication service. Usually the configuration is performed only once and doing it with the help of a graphical interface is easier and more efficient than by using voice or touch-tones.
The notion of Unified Messaging, that is Web and voice access to all messages conveyed in a numerical form: voicemail, e-mails, and faxes, is another example of complementarity of both networks that we have to include in our concept of Personal IP Telephony.

Objectives

The goal of this project is to define the concept of the next generation of personal telephony services and to implement the key notions as part of well chosen prototypes. Since it is our belief that most of traditional IN value-added services will be integrated into the IP world and enhanced by advanced speech processing techniques, the following notions will be included in experimental applications:
- *Voice over IP integration* – A phone-to-phone over Internet is one basic module that could be activated upon request in order to offer low cost, reduced quality communication capabilities. Speech Processing Gateways (SPG) are extensions of Internet telephony gateways including SR, SV and TTS capabilities. They are central to provide distributed and advanced communication services.
- *Inbound call assistance* – Many functions are grouped under this notion: Voice mail, call diversion, who's calling, call waiting, reminder. All of them will be addressed in the concept definition and most of them will be part of Personal IP Telephony prototype.
- *Outbound call assistance* – Voice dialing, information services (e.g. weather forecast), Web-to-phone, least cost routing are among the current foreseen functions a user should profit in the context of an efficient outgoing call completion action.
- *Web control & voice interface* – Control and configuration of personal settings could be greatly improved by a Web access to some parameters, especially if these actions are only rarely ex-

ecuted. A Web site of the EP9701 activities taking advantage of the full range of multimedia possibilities we are mastering, will also be created.
– *Unified Messaging* – Typical functions that should be included here are Internet access to voice mail, e-mail voice delivery.

The Place of academic Research in the EP9701

Past experience has clearly highlighted the benefit Swisscom can draw from collaborating extensively with academic research institutes. We consider the following reasons, among others, as very strategic to play our role of useful intermediary between technology providers and BUs.
First of all, these institutes are implicitly performing technology watch that is of very great importance since time-to-market of a new technology in the speech processing field is getting shorter now than it was only 2 or 3 years ago.
Secondly, it is usually accepted that all the above-mentioned problems are not completely solved for any usage condition. Even if restrictions are known to scientists working in these fields, they cannot be easily understood by people from outside the speech community, and using a system subject to these restrictions may be frustrating for laymen. We consider it important to reduce the scope of some known limitations by addressing very specific problems through projects assigned to research laboratories working in close collaboration with us.
A third reason to actively collaborate with academic research institutes is to have an opportunity to compare state-of-the-art academic solutions with commercially available products. One possible and significant comparison measure is the recognition rate of a recognition system. This rate can be measured on databases of telephone speech samples which are representative of one of the linguistic regions of Switzerland. For speech synthesis, several objective and subjective tests can be carried out in order to assess various technologies, including academic prototypes.
The last but not least, to gain know-how on the current trends in these different research fields in order to reduce the integration time of new solutions with improved results into current and future services.

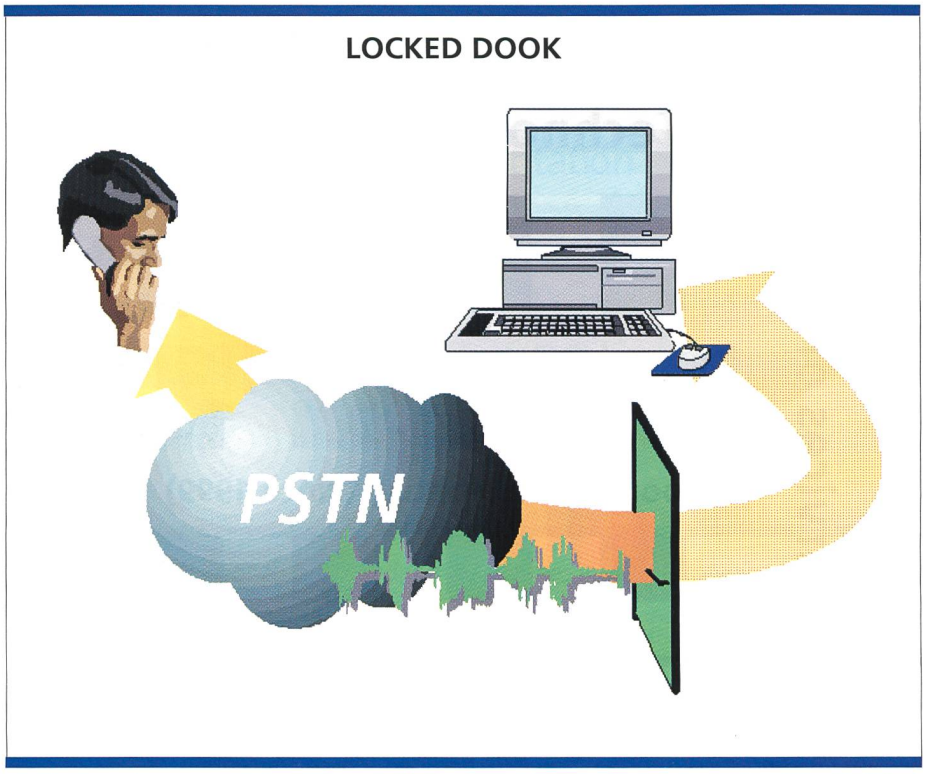

Figure 2. A speaker verification component can be seen as a locked dook, stopping the communication flow between a user and an automatic service. The purpose of this component is clearly to restrict the access to some personal and confidential information or to control access to a paying service. One particularity of this door is that the key to open it is a set of acoustic features specific to the user's voice. This component can thus filter out unauthorized users without any need of a secret password.

The Activities of TIK-ETHZ

The beginning of the activities in speech processing at the ETH dates back to 1976. The first projects were about speech coding and noise cancellation. In the early eighties, two further projects began in speech synthesis and in speaker verification, the latter being completed in the meantime. For the last few years, the activities have been focused on text-to-speech synthesis and speech recognition.
In the course of time, the size of the group varied between five and ten members. Typically, the group members are young postgraduate researchers (electrical engineers, computer scientists and linguists) doing their Ph.D. thesis.

Text-to-speech Synthesis

The ongoing interdisciplinary speech synthesis research is aiming at suitable solutions of all the related problems. Based on these preliminary solutions, a text-to-speech synthesis system for German called SVOX has been developed in collaboration with Swisscom.
Although the speech quality of SVOX is very good, it is currently not suitable in many potential applications. The reason is mainly the absolute restriction to the German language. This makes it completely impossible to synthesize e.g. French or English proper names that are very frequent in every kind of information.
Therefore, the aim of the ongoing research work is primarily to overcome this limitation by incorporating some multilingual capabilities. With this multilinguality, the SVOX system will be able to appropriately articulate German sentences with embedded French, Italian or English words or short expressions.

Continuous Speech Recognition

Continuous speech recognition has to consider the fact that in general a speech signal cannot be transformed into the correct sequence of phones or even characters by means of signal processing methods only. The appropriate way to overcome this problem is the integration of linguistic knowledge into

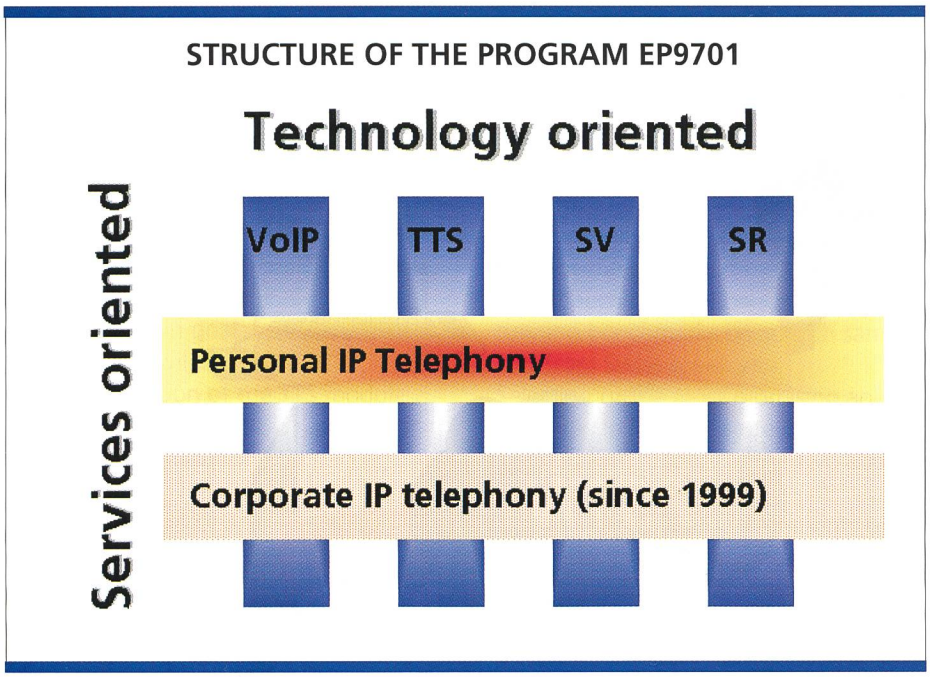

Figure 3. Overall structure of the Exploration Program EP9701, based on four technology research fields, namely Voice over IP (VoIP), text-to-speech (TTS), speaker verification (SV) and speech recognition (SR), and two service concepts.

the recognition process. The ARCOS-G project aims at a speech recognition architecture that closely integrates the signal processing and the linguistic processing components.

The Activities of IDIAP

The Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP) is a semi-private non-profit research institute affiliated with the Swiss Federal Institute of Technology in Lausanne (EPFL) and the University of Geneva. IDIAP is primarily funded by long-term support from the Swiss Confederation, the Canton of Valais, the City of Martigny, and Swisscom. In addition, IDIAP receives substantial research grants from the Swiss National Science Foundation (SNSF) for basic research projects and from the Federal Office for Education and Science (FOES) for European projects.
Focusing on a few well defined research axes, IDIAP carries out fundamental research and develops prototype systems (to validate its research results) along three complementary research directions: speech processing, computer vision and machine learning. For the last few years, IDIAP has counted an average of about 25–30 scientists in residence at IDIAP including permanent staff, postdoctoral fellows, PhD students, and short-medium term visitors.

Continuous Speech Recognition over the Telephone Line

One of the main activities related to the project with Swisscom involves further research and development of speaker independent continuous speech recognition systems over the telephone line. In the telecommunication voice-activated services, speech recognition robustness remains a critical success factor. IDIAP has shown an outstanding knowledge in the field of speech recognition especially within hybrid (HMM/ANN) systems. One of the most promising new trends in Speech Recognition is based on the multi-stream approach. Besides the fact that this technology allows the combination and synchronization of various streams of features, it should, in theory, be less sensitive to noisy speech signals. This work should soon be extended to GSM data.

Speaker verification over the telephone line

Speaker verification at IDIAP is mainly concerned with the improvement of the current state-of-the-art algorithms and the development of innovative solutions combining concurrent and/or complementary strategies. This work is mainly carried out in the framework of one national and two European projects. In the framework of its collaboration with Swisscom, IDIAP developed a Calling Card prototype with speaker verification capabilities.

Management and distribution of speech databases

As part of their collaboration with Swisscom, IDIAP is thus involved in the development (recording and annotation), management, and distribution (e.g., to ELRA) of Swisscom specific large databases of speech samples required to test research results and to develop application prototypes. These databases include the Swiss French Polyphone data (collected by IDIAP) and the Swiss German Polyphone data (collected by ETH, in collaboration with IDIAP). In both cases, the databases consist of the recording of a large number of speakers calling a voice server and being invited to pronounce specific words and sentences, as well as unrestricted requests.

The Activities of LTS-EPFL

The Speech Communication Group (SCG) of the LTS-EPFL (Laboratoire de Traitement des Signaux) was created by Dr. A. Drygajlo and Prof. F. de Coulon in 1993. It is engaged in teaching and research in those aspects of communication engineering that deal with the automatic processing of speech. From early activities including digital spectral analysis and multiresolution processing the SCG has been oriented towards research in speech analysis, compression, synthesis and recognition where the center of activities is speech processing for communication systems.
Currently, the SCG (and its 4 postgraduate researchers doing their Ph.D. thesis) is working on several projects concerning these subjects. The projects are supported not only by the EPFL but also by the Swiss National Science Foundation (SNSF), the Commission for Technology and Innovation (CTI), the Swiss Federal Office for Education and Science (OFES) and industrial partners such as Swisscom and Sun Microsystems (Switzerland).

Multiresolution speech processing in adverse environment for communication systems

The SCG is mostly concerned with speech processing in an adverse environment for voice communication systems. The principal aim of this group of projects is to develop efficient algorithms for speech acquisition, enhancement, com-

pression and recognition which are intended to operate under real conditions. Methods for extracting the perceptually significant speech components from real acoustic signals produced in adverse environments are investigated. These basic research projects with many potential applications attempt to integrate multisensor-, multiresolution time-spectral- and human perception modeling into realistic man-man and man-machine interfaces for robust speech communication systems.

Speech recognition over the telephone (ISDN, GSM)
This group of projects considers applications of robust automatic speech recognition and speaker verification techniques in developing efficient and flexible Interactive Voice Servers (IVS) for Integrated Services Digital Network (ISDN) and mobile (GSM) communication systems. The goal of these projects, supported by the Commission for Technology and Innovation, and industrial partners, is to make available basic technologies for automatic speech recognition and speaker verification on multi-processor SunSPARC workstation and ISDN-GSM platform to industrial partners and particularly to Swiss industry using the Swiss French language.

Speaker recognition over the telephone
The research in this group of projects addresses the problem of automatic multilingual speaker recognition over the telephone lines for communication and forensic application purposes. The issue of integrated methods of speech analysis, feature extraction and model compensation for robust speaker recognition in noisy conditions has become increasingly important since the application of speech/speaker recognition technology in realistic noisy environments becomes an attainable goal. The objective of these projects is to investigate speech processing systems which effectively combine the multiresolution speech decomposition, noise immunity tasks and missing feature compensation.

The Activities of LAIP-UniL

The LAIP was founded 1991 in conjunction with a new chair in computer science for the humanities at the University of Lausanne, and was quickly oriented towards research in speech synthesis. Currently, four experienced speech synthesis researchers are affiliated with our lab, and several more (this year, four contractuals and two interns) are responsible for various specific problem domains. We are involved in several international collaborative projects.

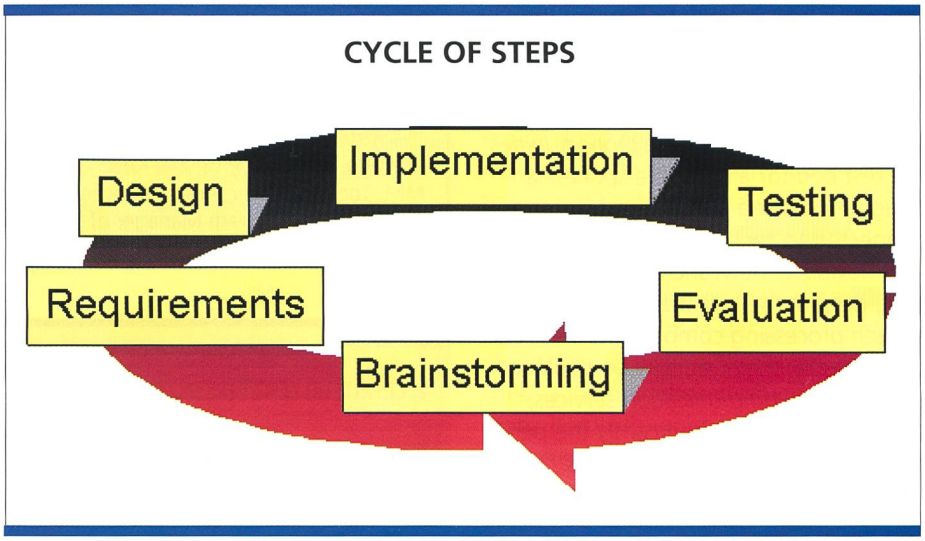

Figure 4. Cycle of steps to be carried out many times in the prototyping of the service-oriented projects.

Like Human Speech
From the beginning, our research was inspired by the principle that improvements in speech synthesis techniques should be guided by an understanding of human speech production and speech perception. We argued that synthesized speech should sound much like humans do. Only that way could we move decidedly beyond the robotic type of speech synthesis that characterized the 1980s and the early 1990s.
Consequently, we systematically examined ways to improve speech synthesis by a direct modeling of human speech. For example, we've implemented a realistic model of word grouping that was inspired by work in psycholinguistics. It produces word groupings that are very much like those that human speakers produce. When words are grouped in expected ways, human listeners find synthesized messages easier to understand. Recently, we've found that these rules change from normal to fast speech. Speakers not only reduce the duration of speech sounds, they also drop some sounds, change some others, and change their pausing behavior. This has direct implications for how we build speech synthesis devices. For example, people with strong visual impairment or blindness use synthesis frequently and are used to listening to very fast speech. To make synthesis optimally adapted for these populations, we must implement different duration mechanisms than for normal speech.
The same argument also goes in the other direction. One of the hardest things to do for a speech synthesizer is to provide address information, especially via the telephone. Here, a clear articulation of each individual sound counts. This is different from giving a weather report, for example. Once the weather forecaster has said "Hier ist der Wetter...", the Swiss German listener expects the word "...bericht". Or in French, once you've heard the words "Voici les prévisions...", the word "météréologiques" is fairly inevitable. This is called "redundancy", and it facilitates the understanding of synthesized speech. In addresses, there is very little redundancy, and as a result, synthesis must imitate human speech even more precisely. If we hope to use speech synthesis for a real telephony application without many customer complaints, every detail of sound duration and corresponding intonation change must be modeled with exceptional precision. Over the next two years, our lab will look at this very question in detail.

Conclusion

The aim of this paper was to give an overview of the activities currently running in the framework of the Exploration Program EP9701. Since part of our activities are carried out in close collaboration with private and academic research insti-

tutions, it was decided to broaden the scope of this presentation by including descriptions of the research interests of these labs.
The EP9701: *Voice Service Opportunities* is a test bed for all possible spoken language processing components tightly connected to capabilities offered by the very flexible IP network. It is our belief that, despite the fact that none of the speech processing components are working perfectly, they can now be deployed in advanced communication services, thereby increasing significantly the performance of value-added services. [4]

Jean-Luc Cochard *received a Dr ès Sc. degree in Computer Science from the Ecole Polytechnique Fédérale de Lausanne in 1988. He then spent two years postdoc at the Canadian Workplace Automation Research Center, Montréal, Canada, and one year at the Istituto Dalle Molle di Studi sull'Intelligenza Artificiale, Lugano, Switzerland. Then he joined the Institut Dalle Molle d'Intelligence Artificielle Perceptive (IDIAP), Martigny, Switzerland, where he was assistant director and research director of the Speech Processing Group. He is now with Swisscom since Nov. 1997. He is working in the field of "automatic speech processing" and plays the role of deputy of the EP9701.*

Acknowledgements

The author wants to acknowledge the contributions made by the following persons to the writing of this paper: Dr. Beat Pfister, TIK-ETHZ, Prof. Hervé Bourlard, IDIAP, Dr. Andrzej Drygajlo, LTS-EPFL, Prof. Eric Keller, Uni, Lausanne, and, on the Swisscom side, MM. Thomas Moser, Aron P. Mueller, Karim Nedir, John Riordan, Paul Vörös, and Robert van Kommer (Program Manager of the EP9701).

Zusammenfassung

Sprachverarbeitung in Telekom-Diensten

Ziel dieses Artikels ist es, einen allgemeinen Überblick der Aktivitäten zu geben, die im Rahmen des Exploration Program EP9701: «Voice Services Opportunities» geplant sind. Da ein wichtiger Teil dieses Programms der automatischen Verarbeitung gesprochener Sprache gewidmet ist, halten wir es für nützlich, mit einer grundlegenden Darstellung dieses Begriffs zu beginnen. Der zentrale Teil dieses Artikels besteht aus einer Beschreibung der Struktur, die wir für die Projekte übernommen haben, welche innerhalb des EP9701 laufen. Neben der Tatsache, dass Forschung betrieben werden muss, um umfassendes Know-how in Sprachverarbeitung und IP-Telefonie zu erwerben, betrachten wir die Schaffung eines starken Geistes der Zusammenarbeit als bedeutendes Anliegen. Wir haben auch beschlossen, die Gelegenheit zu ergreifen, in diesem Artikel die akademischen und privaten Forschungsinstitute vorzustellen, die mit uns für dieses EP9701 zusammenarbeiten. Sie wurden gebeten, ihre Forschungsschwerpunkte kurz zu beschreiben und einige ihrer gegenwärtigen Herausforderungen hervorzuheben. Wir hoffen, dass beide Beschreibungen – applikationsbezogene mit den konkreten Zielsetzungen des EP9701 einerseits und jene mehr grundlagenforschungsorientierten mit den Vorstellungen der Sprachlabors andrerseits – zum Verständnis beitragen, wie komplex die Probleme immer noch sind und was zurzeit auf dem Markt erhältlich ist, um in kurzer Zeit in fortgeschrittenen Kommunikationsdiensten eingeführt zu werden.

COMPUTER '98 in Lausanne ein Erfolg für Swisscom

40 000 Computer-Interessierte besuchten vom 28. April bis 1. Mai 1998 die COMPUTER '98 in Lausanne. Auf einer Gesamtfläche von 25 000 m² vermittelten 429 Anbieter einen Überblick über die Westschweizer Informatiklandschaft.

Die Swisscom AG präsentierte sich als innovativer und kompetenter Lösungsanbieter von Internet, Intranet/Extranet, Security sowie Data/Voice Integration. Alle Swisscom-Lösungen wurden während einer zehn-